Generalizing Concepts and Methods of Verification, Validation, and Accreditation (VV&A) for Military Simulations

Paul K. Davis

RAND
NATIONAL DEFENSE
RESEARCH INSTITUTE

The research described in this report was sponsored by the Under Secretary of Defense for Acquisition. The research was conducted in the National Defense Research Institute, RAND's federally funded research and development center supported by the Office of the Secretary of Defense and the Joint Staff, Contract No. MDA903-90-C-0004.

ISBN: 0-8330-1298-3

Published 1992 by RAND
1700 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138

R-4249-ACQ

Generalizing Concepts and Methods of Verification, Validation, and Accreditation (VV&A) for Military Simulations

Paul K. Davis

Prepared for the
Under Secretary of Defense for Acquisition

RAND

PREFACE

This study was developed for the Defense Modeling and Simulation Office (DMSO), which is under the Director, Defense Research and Engineering. It reflects discussions of the DMSO's Applications and Methodology Working Group, chaired by the author during this work. The study also draws upon discussions at two special meetings on verification, validation, and accreditation (VV&A) sponsored by the Military Operations Research Society (MORS) on October 15–18, 1990, and March 31–April 2, 1992. VV&A is a difficult subject on which there is a broad range of opinions and practices (for example, VV&A of software used in space probes is different from VV&A of military simulations used for analysis). At the same time, a considerable convergence of view appears to be taking place and it is hoped that this study will accelerate that process. Comments and suggestions are therefore especially welcome. They can be sent by electronic mail to Paul_Davis@rand.org through Inter Net.

Work on this effort was accomplished in the Applied Science and Technology program of RAND's National Defense Research Institute, a federally funded research and development center sponsored by the Office of the Secretary of Defense and the Joint Staff.

SUMMARY

This study on verification, validation, and accreditation (VV&A) seeks, for military models and simulations: (a) to provide a simple and realistic framework for modelers, analysts, managers, and recipients of analysis, (b) to address important complications that have received too little attention in the past (for example, evaluation of knowledge-based models such as those representing command and control decisions and other behaviors), and (c) to discuss how modern model-building technology is changing the way we *should* develop models and conduct VV&A. The study illustrates many of its suggestions about VV&A with down-to-earth examples of language that might be used in reports and accreditation reviews. It sketches elements of advanced modeling and analysis environments that would make such work easier.

A VV&A FRAMEWORK THAT REFLECTS MODEL DIVERSITY AND FOCUSES ON APPLICATIONS

Military models and simulations vary substantially in character and basis. Some can be evaluated by direct comparison with physical measurements. Others can at best be evaluated by establishing that they reflect accurately the judgments of particular experts. Some models represent well the best and most reliable information available, but other models representing the best information available are not reliable because the underlying phenomena are poorly understood. Some models are excellent for predicting expected-value results, but the phenomena in question are stochastic. And so on. Because of such diversity, concepts and managerial regimes for VV&A should apply a broad range of evaluation methods and should recognize that conclusions about validity are likely in most cases to remain provisional and conditional. Further, it should be emphasized at every opportunity that judgments about model validity can usually be made intelligently only in the context of a specific application and should include a statement about subjective confidence (or, equivalently, about residual uncertainties of all types). Judgments of validity should often depend not only on the *type* of application, but on the detailed manner in which model outputs are used to reach conclusions or characterize alternatives.

With these considerations in mind, I recommend: (a) a new definition of *generalized validation (evaluation)* that highlights different dimen-

sions of validity and the important issue of subjective confidence, (b) a taxonomy of VV&A (Figure S.1) that focuses on validation and includes special methods for knowledge-based models, and (c) an application-centered process model (Figure S.2) of how VV&A should be conducted. In this approach, VV&A is seen as an iterative process that may extend over years, especially with successful models used in a series of applications. Figure S.1 can be used to guide planning by providing a checklist of methods to be used over a period of months or years (for example, using both historical data and conducting special field tests).

The approach I recommend to accreditation is unusual. In my view, accreditation should be a *substantive* process in which those overseeing a particular application provide guidance on the appropriateness of the analytic plan for using models (note the output on the bottom right of Figure S.2), guidance that may include clear-cut instructions about the types of conclusions that should and should not be attempted and about specific sensitivity analyses that must be presented to assure that analysis recipients appreciate the most relevant uncertainties. This accreditation decision should also reflect judgments about the team of people conducting the application and about the comprehensibility of the analysis. Accrediting authorities should prefer and encourage comprehensible models with explanation capabilities and without excessive detail.

Another virtue of the process view shown in Figure S.2 is that it highlights the need to assess the value of VV&A activities in terms of accomplishing the mission (e.g., an analysis). There are costs associated with the special model adaptations and special verification and validation tests indicated in the figure; some are more worthwhile than others. If sponsors of the effort are unwilling or unable to fund VV&A adequately, then the scope of the analysis should be scaled back or the claims made about the analysis restrained. VV&A can be rather costly (it can account for 20 percent of model-development costs) and time consuming (it may take 6 months for serious independent verification testing). On the other hand, VV&A is very important and has long been inadequately funded by any measure. By explicitly budgeting for "serious" VV&A, the Department of Defense would create incentives that do not now exist for model developers. Without such incentives, VV&A may improve only marginally, despite the suggestions and exhortations from this and other studies.

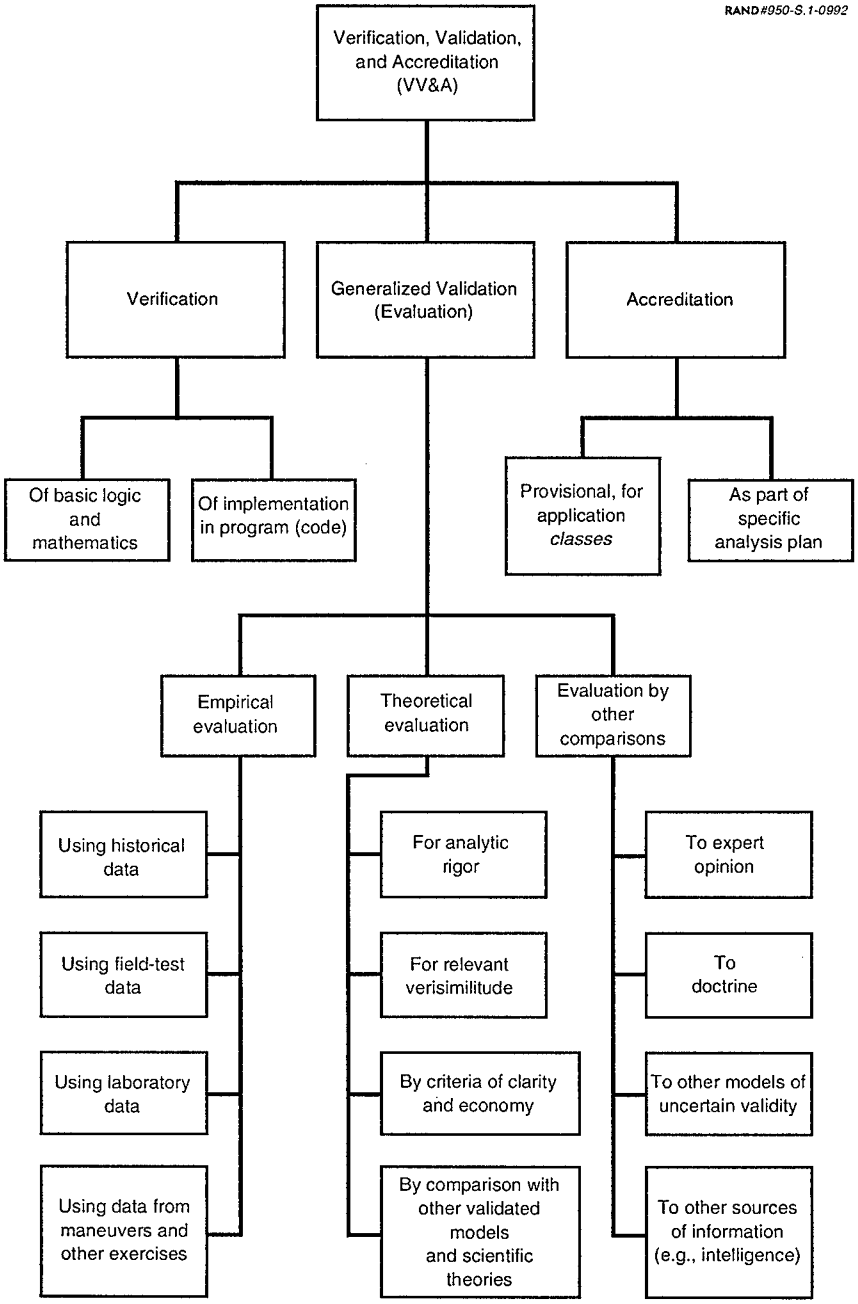
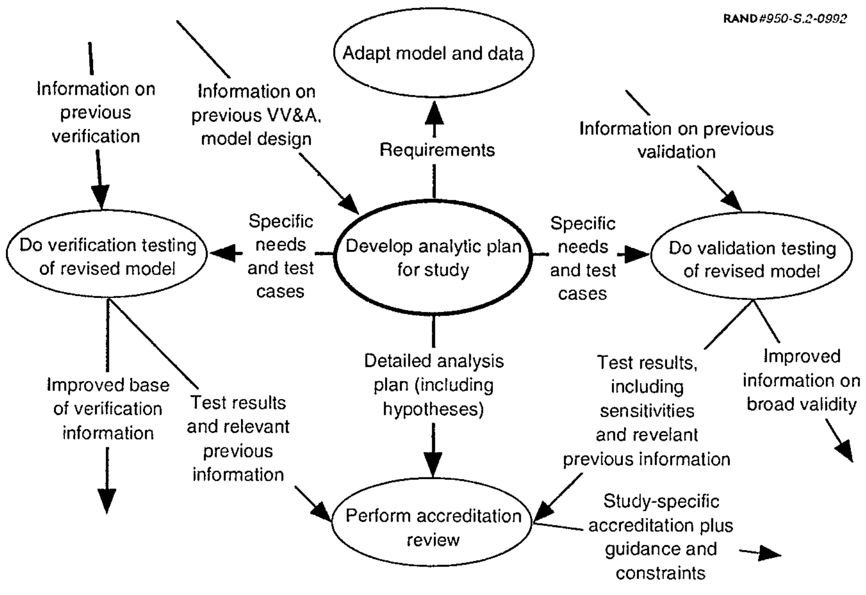

RAND#950-S.1-0992
Verification, Validation, and Accreditation (VV&A)
Verification
Generalized Validation (Evaluation)
Accreditation
Of basic logic and mathematics
Of implementation in program (code)
Provisional, for application *classes*
As part of specific analysis plan
Empirical evaluation
Theoretical evaluation
Evaluation by other comparisons
Using historical data
Using field-test data
Using laboratory data
Using data from maneuvers and other exercises
For analytic rigor
For relevant verisimilitude
By criteria of clarity and economy
By comparison with other validated models and scientific theories
To expert opinion
To doctrine
To other models of uncertain validity
To other sources of information (e.g., intelligence)

Figure S.1—A Taxonomy for VV&A


RAND#950-S.2-0992
Adapt model and data
Information on previous verification
Information on previous VV&A, model design
Requirements
Information on previous validation
Do verification testing of revised model
Specific needs and test cases
Develop analytic plan for study
Specific needs and test cases
Do validation testing of revised model
Improved base of verification information
Test results and relevant previous information
Detailed analysis plan (including hypotheses)
Test results, including sensitivities and revelant previous information
Improved information on broad validity
Perform accreditation review
Study-specific accreditation plus guidance and constraints

Most processes are iterative (not shown)

Figure S.2—An Applications-Centered Dynamic View of VV&A
(start at center, assuming understanding of problem being addressed)

IMPLICATIONS OF MODERN TECHNOLOGY

Many of the prior articles about VV&A were written before recent changes of technology that should be changing the methods used for model development generally, including VV&A. For example, the traditional recommendation by scholars to separate development and evaluation of the conceptual model from development and evaluation of the computer program is no longer as appropriate as it once was. A rethinking of this entire issue is needed as technology in the form of high-level languages, advanced modeling and analysis environments, and rapid prototyping methods blurs the distinctions between model and program in ways that are, on balance, beneficial. New methods are needed for design, parallel documentation, automated verification tests, automated explanation of model results, and many other functions. If the DoD community is to exploit these emerging opportunities, which could improve model development and VV&A greatly, the DoD should invest more heavily than it has in recent years in developing modern modeling and analysis "environments." Since the best work on such matters depends on exploiting commercial tools (e.g., for graphics, spreadsheets, hypertext, and object-oriented programming), and common languages and operating systems such as C/Unix, the DoD should avoid requiring modeling efforts to adopt a particular format and language such as Ada. High degrees of reusability and interoperability can be achieved by establishing standards that are not language specific. Indeed, that is what makes the increasingly common emphasis on "open architectures" feasible.

ACKNOWLEDGMENTS

This study benefited from numerous discussions with colleagues in the analytic community, including members of the DMSO's Applications and Methodology Working Group and a special panel of the Military Operations Research Society concerned with VV&A. Nonetheless, the material presented here is my responsibility and I make no claims about consensus in the community. I particularly appreciate formal reviews by Dennis Shea (Center for Naval Analyses) and Richard Hillestad (RAND), and the many discussions with or informal review comments provided by: Clayton Thomas (Air Force Center for Studies and Analysis); Dale Henderson (Los Alamos National Laboratory); COL. Gilbert Brauch (Army Model and Simulation Management Office); Marion Williams (Air Force Operational Testing and Evaluation Command); James Sikora (BDM, Inc.); Richard Bronowitz (Center for Naval Analyses); Julian Palmore (University of Illinois); and RAND colleagues James Dewar, James Hodges, Richard Hundley, Iris Kameny, Eric Larson, and Clairice Veit. Residual problems are, of course, my responsibility.

CONTENTS

FIGURES

TABLE

1. INTRODUCTION

OBJECTIVES

Verification, validation, and accreditation (VV&A) is a complex subject that has troubled model developers and users for many years. Each generation of modelers and analysts must think it through, because understanding the issues is important to professionalism. Consumers of analyses exploiting models must also understand the subject or they will have difficulty judging the quality of products. Further, they may be either insufficiently demanding or supportive of VV&A efforts on one extreme, or unreasonable on the other—requiring a degree of validation that is impossible even in principle. Managers of analysis organizations should understand VV&A so that they can put into place appropriate procedures, standards, and incentives. This may be called a VV&A "regime" to emphasize that VV&A is not a one-time event, but rather an ongoing but episodic *organizational* activity that should be understood and considered important by all participants.

What, then, might a VV&A regime look like if one saw it? What advice should be given to a new manager who is ready and willing to institute reforms to establish sound VV&A policies and procedures? This study attempts to sketch the essential features of an answer. Its principal objective is to provide guidance that would be useful to such a manager in government, industry, or the academic world. Auxiliary objectives include discussing the special VV&A problems associated with knowledge-based models and recommending new attitudes about model development and VV&A that reflect the implications of modern technology.

BACKGROUND

There is a considerable literature on VV&A for military models, much of it severely critical of model developers and their government sponsors for there not having been enough VV&A in the past.[1] There is no

[1] Standard references of this sort include Shubik and Brewer (1972), U.S. GAO (1980), and U.S. GAO (1987), which contains an extensive bibliography. One of the most famous essays on the subject is Stockfisch (1973). Davis and Blumenthal (1991) examine broader issues and argue that many problems in combat modeling stem from failure of the military community to think in terms of nurturing a robust military *science*. The near-exclusive emphasis on models as mere tools has been an obstacle to

definitive source on what VV&A is or should be, but someone new to the field might well consult Thomas (1983), other chapters of Hughes (1989), Gass (1983), Sargent (1987), and Martin Marrietta (1990).[2] The first of these has a philosophical slant and addresses some of the profound difficulties in even contemplating model evaluation. The latter, which draws on the work of Gass, Sargent, and others, describes an approach that has been used in large-scale efforts having to pass rather stringent DoD criteria. Another good introduction to validation issues is Miser and Quade (1988). Finally, those concerned with VV&A will surely want to examine guideline documents emerging from sponsoring organizations, as well as regulatory documents such as U.S. Army (1992) (especially Chapter 6 on VV&A) and DoD-MIL-STD 2167, which describes software standards.

In this study I present some definitions (Section 2) and discuss what the definitions mean and why they are not simpler. My definitions of validation and accreditation extend the more usual ones in important ways. Section 2 then presents a taxonomy of VV&A methods, focusing primarily on validation. Section 3 describes VV&A as a dynamic *process* that should conduct evaluations both for broad classes of model application and for specific studies having detailed analytic plans. Section 4 then pulls things together and recommends an approach for the use of practitioners, managers, and consumers of model-based analysis.

seeing some models as *theories* that need to be developed, tested, and evolved scientifically.

[2] Another useful reference is Williams and Sikora (1991), which provides a snapshot view of continuing work on VV&A by the Military Operations Research Society (MORS). Readers may wish to check for updates in the newsletter *Phalanx*. MORS hopes to publish a book on VV&A in 1993. This study may contribute to that effort.

2. DEFINITIONS AND CONCEPTS

MODELS AND PROGRAMS

"Models" are representations of certain aspects of reality (for example, of certain aspects of particular systems). They come in many forms, including the physical scale models used by architects, analytical models expressed in paper-and-pencil equations, and computer models (see also the overview chapter of Hughes, 1989). In this study, I concentrate on computerized models, primarily "simulation models," which attempt to describe how a system changes (behaves) over time.[1] I am also concerned here with models having phenomenological content relating causes and effects rather than, say, regression "models" or optimizing algorithms that some might call models.

Although the terms "model," "simulation," and "program" are often used interchangeably, here and elsewhere, it is sometimes important to make distinctions, especially between the model (or what some call the conceptual model) and the program (or computer code) that implements the model. Appendix A elaborates on this and argues, reluctantly and in contradiction with the advice given by most scholars, that it is becoming increasingly difficult—and decreasingly appropriate—to separate the processes of designing and evaluating models on the one hand, and designing, building, and evaluating program implementations on the other. Technological change demands a new approach.

MODELS, DATA, AND KNOWLEDGE BASES

Throughout this study "model" means the union of a "bare model" (also referred to as "the model itself") and its "data base." Thus, $Y(t) = Y(0) - 1/2\ g\ t^2$ is a bare model, whereas $\{g = 32\ \text{ft/sec}^2;\ Y(0) = 10{,}000\ \text{ft}\}$ is a data base. In some instances, the data represent a "knowledge base" in the form of rules and algorithms.

In the past, bare models were conceptually distinct from data in most cases. The bare models defined structure and algorithms; the data base provided values (for the gravitational constant or the number of tanks in a division, for example). Modern practice, however, has

[1] Some sources define "simulation" differently—as the operation or exercise of a model, or as a method of implementation.

muddied the distinctions. In many models, much of the substantive content is defined in the data base because with most computer models it is easier and faster to change data than the program itself and developers have sought to provide users as much flexibility as possible.[2] As a result, *the VV&A process must consider both bare models and data bases.*[3] Quite often, bare models and data bases need to be reviewed together, in the context of an application; in other cases (i.e., with different model designs), they can to greater or lesser degree be reviewed separately. For example, one can conduct VV&A on an order-of-battle data base without knowing precisely how that data base will be used. Similarly, one can conduct VV&A on an algorithm without knowing the precise context in which it will be used.

VERIFICATION

Verification is the process of determining that a model implementation (i.e., a program) accurately represents the developer's conceptual description and specifications.

This is the definition commonly accepted in the military modeling community. There continues, however, to be some confusion and disagreement about precisely what is and is not covered under verification, and about what taxonomy describes verification activities. I consider verification to consist of two basic parts.

- *Logical and mathematical verification* ensures that the basic algorithms and rules are as intended by the designer and do not include logical or mathematical errors (e.g., divisions by zero, incompletely specified logic, or nonsense results when certain variables

[2] As an example, consider a model predicting the damage expectancy for a set of hard targets as a function of a bomber's availability, reliability, pre-launch survivability, penetration probability, bomb load, and hard-target kill capability. The bare model provides an intellectual framework, but has little or no predictive value: Its predictions are "data driven." Similarly, in idealized knowledge-based systems such as an expert system describing likely decisions of a commander, the bare model may be a general "inference engine" for processing rules, while the content of the model resides entirely in the "knowledge base" of rules such as "If we can achieve surprise and if the force ratio is no worse than . . . Then we shall . . ."

[3] The introduction of highly interactive computer languages has broken down the classical distinction between model and data, which makes it possible for users to change many equations and structures in the computer code as easily as they can change the data value used for the gravitational constant. The most familiar example of this is in spreadsheet programs, but other examples include BASIC and RAND-ABEL®. (RAND-ABEL is a trademark of RAND.)

take extreme or unusual values). Although verification is nominally concerned with implementation rather than correctness of design, it is common for verification activities to uncover design errors along the way (for example, to detect an implicit and unreasonable assumption about independence of events). Thus, verification activities should begin with documentation and will often accomplish some validation functions.

- *Program verification* (*or code verification*) ensures that these representations have been correctly implemented in the computer program. Program verification is concerned in part with simple matters such as discovering and correcting typographical errors, errors in the units in which physical quantities are described, and errors of definition (for example, a model designer might have intended that a force ratio apply only to forces on the forward line of own troops (FLOT), but the programmer might have defined it to apply to groupings that include corps-level reserves). It is also concerned with more complex issues such as the appropriateness of numerical integration techniques,[4] covering all the logical cases (including cases that the designer might consider unlikely or unphysical), and eliminating bugs that would cause the program to "crash" in some circumstances. Many such bugs involve intricacies of the particular computer hardware, operating system, and interface software.

Verification is a matter of degree for complex models, because it is impossible in practice to test the model over the entire range of variable values and because it is often not feasible with available resources to do a line-by-line code check. Thus, a model may be well verified within a particular "scenario space," but not well verified otherwise.[5] In principle, one might think of using sampling techniques to verify code to some level of confidence, but I am personally unaware of any rigorous efforts to do so in the realm of combat modeling.

Verification of *data* (especially classical types of data such as physical constants or orders of battle rather than, say, data defining elements of model structure or exponents in algorithms) should often be distinguished from verification of the bare model, because different tech-

[4] A related issue here is establishing that the numerical procedures used are not introducing chaos effects. See, for example, Palmore (1992).

[5] Articles on software engineering sometimes use terms such as "rigorous audit" or otherwise convey the impression of verification requiring complete testing over all computational "paths." Except at the level of relatively small modules, however, such review and testing is usually not feasible. Thus, there is a premium on designing a doable set of tests that will be likely to uncover the most serious problems.

niques are involved and data bases change frequently.[6] There are at least two aspects of data verification. The first aspect involves ensuring that source data are converted properly to model input data and are consistent with the model concept and logical design (for example, that data supposed to represent conditional probabilities of kill given a hit do indeed represent those rather than, say, kill probabilities per *shot*). It should also include spot checks to confirm that data were, in fact, extracted from the stated source and that they represent the latest available from that source. If data are not provided with the model, then verification should include establishing that the required user inputs are readily available.

A different aspect of data verification applies within the context of a study if the data base has already been installed. Here one seeks to establish whether the data base represents correctly the assumptions intended for the analysis. For example, if an analyst states that he wants to use a particular official data base for orders of battle, database verification would include checking that the desired data base was the starting point for the installed data base, but it would also check to see if appropriate corrections had been made—corrections that the analyst would surely want if only he knew to ask for them. These would include providing realistic data values where the original data base had zeros, blanks, or values annotated as "purely nominal." Official data bases are often riddled with holes and errors. Managers of analysis and recipients of analysis are often unaware of how serious these holes and errors are, or of how much the analysis depends on the cleaning-up process, which often requires substantive work and numerous subjective judgments (which unavoidably mixes verification and validation activities).[7]

VALIDATION

Validation is the process of determining: (a) the manner in which and degree to which a model (and its data) is an accurate representation of the real world from the perspective of the intended uses of the model and (b) the subjective confidence that should be placed on this assessment.

[6] Some of the following discussion draws on review comments by Dennis Shea of the Center for Naval Analyses. See also Pace and Shea (1992).

[7] As discussed below, a number of modern techniques can automate or otherwise assist a good deal of verification testing. Many depend on the existence of a data dictionary that is part of the language or environment, not a mere repository of comments.

This definition extends the more conventional definition.[8] The extension calls attention to two considerations. First, there are different meanings to "accurate representation." Second, the validation process should address the issue of confidence (not in the sense of "statistical confidence," but in the larger sense having to do with how much one would bet on the correctness of the model's predictions given residual uncertainties). While one could consider both considerations to be implicit in the more usual definition, it seems to me evident from experience that they will be underappreciated unless made explicit.

Types of Validity

To elaborate on the definition given above for "validation," I use the phrase "manner in which" because a model can be "valid" in several distinct ways. It may have (a) descriptive validity, (b) structural validity, or (c) predictive validity (see also Zeigler, 1984).[9]

Descriptive validity means here that the model is able to *explain* phenomena or organize information meaningfully in one way or another. For example, a descriptive model might be able to say, "Well, the reason this happened is that A collided with B, which happened because A had lost its radar and therefore failed to see B in the cloud bank." All of this might be a sound and nontrivial reconstruction of events. Note that the model used for such a reconstruction might not have been able to predict the events ahead of time, especially if the key causative events were stochastic or some key inputs such as precise speed histories were unknown. What constitutes a "good" description or explanation depends on context and taste.

Structural validity means that the model has the appropriate entities (objects), attributes (variables), and processes so that it corresponds in that sense to the real world (verisimilitude), at least as viewed at a particular level of resolution.[10] One may also require, for structural validity, that the principal algorithms are at least roughly appropriate, although not necessarily accurate (e.g., whether a process de-

[8] As of April 1, 1992, the MORS group concerned with VV&A was using as a working definition: "The process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model."

[9] Other workers sometimes refer to structural validity vs output validity. In that breakdown, output validity includes both descriptive and predictive validity.

[10] The subject of resolution is complex and analysts often need to work with models with different resolutions which, ideally, are consistent in the aggregate. See Davis and Huber (1992) for a related discussion.

scribes exponential or linear growth may be regarded as a structural issue).

Predictive validity means that a model (including available or potentially available data) can predict desired features of system behavior, at least for particular domains of the initial conditions and durations of time, to within some known level of accuracy and precision. A conditionally predictive model explicitly identifies alternative behaviors and the conditions that would cause them (e.g., "*If* the weather tomorrow remains clear, *then* the air operation should go well and . . .").

These types of validity can be considered more or less orthogonal attributes of a model. As suggested in Table 2.1, one can have models with every combination of type validity. This is significant to VV&A, because the criteria one applies depend strongly on the type of validity sought.

To illustrate a few points in Table 2.1, consider first that a model can be excellent, even definitive, for explaining phenomena *after the fact*, and yet be useless for prediction (e.g., Case 2), at least in the usual sense. This happens if the model depends on the values of variables that are unknown until after the fact (e.g., the fighting quality of the other sides' forces). This situation occurs commonly with military models, since we do not know the detailed initial conditions for future military operations. Nor do we know the various decisions that will be made in the course of operations. After the fact, these decisions and other previously unknowable variables may be unambiguous and objective data (e.g., as reflected in operations orders and reports on what the weather was). If the model then explains the phenomena well in retrospect (sensibly as well as accurately), the model is descriptive.[11]

As a second example, structural validity does *not* imply that the attribute values are correct or that the algorithms constituting the model processes are precise. A model of combat might be structurally valid while treating attrition quite approximately: It would have an attrition process, but the process would be inaccurate (Case 6).

The most subtle example here is probably that predictive validity does not imply descriptive validity, in our sense. One can have an empirically based model, perhaps in statistical form, that is remark-

[11] One effort to assess descriptive validity is described in Bonder (1984), which examines the ability of the Vector-2 model to reproduce the battle for the Golan Heights in the Yom Kippur War.

Table 2.1

Models with Different Combinations of Validity Type

Case	Descriptive Validity	Structural Validity	Predictive Validity	Example
1	Yes	Yes	Yes	Well-tested weapons-performance models.
2	Yes	Yes	No	Good theater-level models (which may, however, be conditionally predictive for some features of a campaign, at least in certain domains such as when one side has overwhelming force).
3	Yes	No	No	Historically based statistical models correlating different measures of outcome (e.g., movement rate and ratio of loss rates; one might say "*Because* the ratio of loss rates was low, the advance rate was fast").
4	Yes	No	Yes	Some highly aggregated models that reflect doctrine and experience (e.g., march times for unopposed moves).
5	No	Yes	Yes	Incomprehensible but reliable black-box models with high resolution in entities and processes (e.g., poorly coded models with little documentation or explanation capability).
6	No	Yes	No	Models with high resolution in entities and processes, but poor algorithms (e.g., weapon-on-weapon attrition calculations assuming perfect tactical command and control).
7	No	No	Yes	Rules-of-thumb models or statistical models that work for no clear reason and do not represent system structure (e.g., a regression model predicting the next week's weather as a function of today's weather and the month of year).
8	No	No	No	Bad models.

ably predictive but that says little or nothing about the cause-effect relationships at the levels of physical entities and processes (Case 7). It is often difficult to know when such models will fail, but they are useful nonetheless.[12,13]

Again, then, the point here is that evaluation of models should vary with type. It is silly to denigrate a good descriptive model that is structurally valid, merely because it is not a prediction machine (given the data known ahead of time). This is nontrivial, because many critics of military modeling are guilty of precisely this error. Those who argue that attrition estimates for the Desert Storm operation were off by an order of magnitude overlook the fact that many analysts were explicit about their estimates being upper bounds and about the potential for much lower attrition if the Iraqis proved ineffective by virtue of poor morale, training, and leadership.

Issues of Degree and Confidence

The words "degree" and "confidence" appear in my definition of "validity," because models are seldom perfectly valid in any of the dimensions (description, structure, or prediction). They vary in their accuracy and precision. Also, *there are several dimensions of confidence*, since:

- The model or its data may be known to be highly uncertain (for example, in functional form or in data values).
- The model and its data may represent a best-estimate consensus of experts, but may nonetheless be fundamentally wrong (e.g., Ptolemaic astronomy). One dimension of confidence, then, relates to assessing the likelihood of the bare model or its data having serious flaws not yet thought of or taken seriously.

[12] Other decompositions are possible. Working from discussions at the MORS SIMVAL II meeting, Dale Henderson of Los Alamos National Laboratory decomposes validation activities into five areas: (a) the techniques used (e.g., Delphi vs quantitative comparisons), (b) the basis of truth used (e.g., historical data vs results of more detailed simulations), (c) the applications intended for the model, (d) the degree of composition at which testing occurs (e.g., on primitive modules vs higher-level subsystems or a complete integrated system), and (e) the depth of the validation effort (e.g., surface-level or face-validity testing). The principal point is that validation activities are multidimensional rather than rank-ordered or hierarchical.

[13] Prehistoric man presumably "knew" that the sun would come up every morning and that there was a cycle of progressively longer and then progressively shorter days. He presumably counted on this model long before there was any understanding of astronomy.

- A model may be deterministic, while the relevant world may be stochastic.[14] In this case, confidence in the model's predictiveness depends on the underlying probability distributions. If the distribution function is strongly weighted around a central point, then a deterministic model may be reasonable; if the function is bimodal, then such a deterministic model may be downright misleading.

For all of these reasons, the process of validation should include reaching explicit, albeit often subjective, judgments about the confidence one places in the model. These can be aided by sensitivity analyses coupled with analysis assessing how much one *truly* knows about the more critical variables in the context of a shooting war.

Some examples may be useful here to illustrate how central the issue of confidence really is in the use of military models. Consider the following hypothetical statements about models being made by analysts to general officers in the context of a real war or preparations for such a war:

> The strategic-mobility model itself is solid, for aggregate predictions, but predictions depend on planning factors and decisions. We should plan for buildup rates +/- 30 percent around baseline data. Also, we should recognize that the CINC may make significant changes in the Time-Phased Force Deployment List, so we must anticipate the kinds of changes he would most likely seek and consider their consequences on predicted buildup rate.
>
> Because of uncertainties, including random factors and intrabattle decisions, we have no confidence in predicting winner or loser (or low casualties)—unless we can stack the deck by going for a 6:1 local force ratio after bombing. Then we would be confident.
>
> Results will depend on surprise and speed. That's beyond our model's ability to predict well. The model is descriptive after the fact, but that doesn't tell us what we need to know now. We can instead tell you, as a commander, how quickly we think you *need* to maneuver for success, based on intelligence estimates on the enemy's reaction times and maneuver speeds as judged from doctrine and exercises over the last few years. Whether you can do that is difficult for us to judge.
>
> The ECM-ECCM model is very accurate for aircraft flying against the SA-99 as we know it, but the enemy may have changed subsystems, in which case noise jamming would be unchanged but false-target generation might not work at all. We simply don't know whether he has changed systems.

[14] Most of our "stochastic processes" are at their root deterministic; the problem is our uncertainty about initial values and interactions with other processes, which causes us to treat them as stochastic.

All of these statements could be made quantitative to avoid ambiguity, but my recommendation is to use the language of odds in a context that downplays confidence and reminds everyone of the stakes (e.g., mens' lives) rather than using the language and tone of statistical precision. As an example:

> If we have characterized the SA-99 correctly, as we *think* we have, our ECM should be less than 1 percent (between about 0.5 percent and 1 percent). If the enemy has changed subsystems and can defeat our false-target generation (this is highly subjective, but I'd say that's a 1-in-4 situation), then our rough calculations suggest our losses will be about 1–2 percent per sortie until we can destroy the surface-to-air missiles. Even in the bad case, we estimate that losses won't be worse than 3–4 percent per sortie because they have a limited number of SAMs. That loss rate might last up to three or four days, but we're very confident we will destroy the SAMs in no more than that time.

Data Validation

In most of this study, I treat data validation as part of validation generally. It is worth mentioning some unique features of data validation, however. These relate primarily to the types of data one uses to introduce facts, official estimates, and other numbers rather than, say, the types of data one may use to define aspects of the model (for example, spatial resolution or exponents in equations). In this activity, one typically reviews the data sources and how they were collected to compare model input data to real-world or best-estimate values. This may involve assessing the credibility of data sources and comparing alternative data bases. In reviewing operational data, one must consider exercise artificialities such as safety-related constraints and geography. Data validation is often quite troublesome. Intelligence estimates, for example, may vary widely with little rationale given, and estimates of system effectiveness for U.S. weapons are often extrapolations from small data samples collected under artificial conditions.

ACCREDITATION

Accreditation (often used synonymously with certification) is an official determination that a model is acceptable for a specific purpose (e.g., to a class of applications or to a particular analysis or exercise).

Accreditation by Class of Application vs Specific Application

Except for the parenthetical phrases, this is a commonly accepted definition (e.g., Williams and Sikora, 1991, and U.S. Army, 1992). It says that accreditation is a *decision* (not just a process) to the effect that a given level and character of verification and validation are sufficient to justify using a model in a particular application.[15]

Problems arise not with the definition but with what organizations charged with model VV&A sometimes try to do. It would be convenient for such organizations if models could be definitively accredited for broad *classes* of applications, but even within a given class of applications (e.g., weapon-system comparisons), a model will sometimes be adequate and sometimes not. Which situation applies depends on details, including numerical details and the sensitivity of results to errors in model performance. Also, some models that might be thought inappropriate to a particular application can be used effectively if manipulated cleverly with the benefit of parametric variations informed by side calculations.[16] *It follows that class-level accreditation should be provisional only, and that accrediting authorities should be extremely cautious in claiming that models cannot or should not be used for applications within a given class.* Those long familiar with VV&A issues and organizational behavior are perhaps most concerned about this problem, because they see the potential for mischief when controversial studies use models. Another concern here stems from the observation that organizations sometimes insist that "accredited models" be used for studies even when those models are inappropriate compared to alternatives that have not yet been accredited, or even fully developed. Furthermore, many fear that the accreditation process will place too much of a premium on verisimilitude and too little emphasis on clarity, controllability, and efficiency.

[15] In practice, application-specific accreditation usually depends (and *should* depend) on an assessment of the people and organization using the model, not merely the model itself. Indeed, one can argue that it is more important to "accredit" (or at least to assess) people and organizations than the tools they use.

[16] A classic example of this is use of silo hardness, measured in psi. Many strategic-nuclear analyses have been conducted using silo hardness, even though the phenomenology of silo destruction is complex and requires something more sophisticated, such as a vulnerability number approach that accounts for effects of both static and dynamic pressures. Analysts can nonetheless get by with computer programs or analytic models using hardness, because they do offline calculations to derive the effective hardness of silos to the weapon yields of interest.

A Crucial Issue in Sound Accreditation: Model Clarity

It is perhaps a symptom of the disconnect between analysts and those who build and sponsor models that discussions of VV&A seldom mention one of the most important considerations in evaluating a model: its *clarity*. One could argue that the definition of validation should be modified to include such considerations, but I have chosen in this study to argue that these considerations are very much in the province of those who oversee particular uses of models. They have an important stake in model clarity, because:

- They are responsible for results and their ability to review the work (or have it reviewed by independent experts) depends on their ability to comprehend the model and the cause-effect relationships dominating results.
- They are responsible for communicating results, which typically requires separating essentials from noise.
- They may want to be able to reproduce the work, which will be far easier if it has been conducted with a comprehensible model.

It follows, then, that accreditation should depend not only on the soundness of the model for the application at hand, but on the ease with which the model can be comprehended and the results of the model understood in terms of appropriate cause-effect relationships. *That is, model accreditation should depend not only on model soundness for the application, but also on: (a) comprehensibility of the model and (b) comprehensibility of model runs (through "explanation capabilities").* This facet of the problem has been greatly underappreciated in prior discussions of VV&A, even within the academic community and even by systems analysts, who certainly wax eloquent about the need for model simplicity in other contexts. I observe also that the importance of model clarity increases the importance of establishing a model's descriptive validity.[17]

[17] One can argue that the issue of clarity applies more to the *study* or other application than to the model itself, but those interested in the clarity (and reproducibility) of studies are usually driven toward seeking clarity of models as well. While it is true *in principle* that analysis with black-box models can be clear, given enough sensitivity testing, my own experience is that depending on such an approach is usually a recipe for disaster.

3. A TAXONOMIC VIEW: THE CONSTITUENTS OF VV&A

PREFATORY DISTINCTIONS

Given the above definitions, how does one *accomplish* VV&A? Suppose one is attempting to establish a *VV&A regime* within an organization, a regime in which one routinely does virtuous evaluation before using models for analysis. How does one go about it?

It is useful first to make some distinctions:

- Components vs system (or modules vs integrated model)
- Bare models vs data;
- Evaluating "best estimate" functional forms and data values vs evaluating ranges, distributions, and confidence;
- Conducting "broad VV&A" with only a partial sense of the intended applications vs conducting focused VV&A for a particular study.[1]

VV&A applies to each half of each of these pairs. I emphasize this at the start, rather than repeating it at every point of the following discussion.

A STRUCTURAL PERSPECTIVE: THE COMPONENTS OF VV&A

Figure 3.1 now provides a *structural*, or taxonomic, view of what constitutes VV&A. It elaborates on validation, because that aspect has been most controversial and confusing over the years. I use the phrase "generalized validation" or "evaluation" here, because my sense of validation is broader than that of some authors.

VERIFICATION METHODS

Although this study does not emphasize verification methods (see Sargent, 1987, and Martin Marrietta, 1990, for more discussion), the

[1] As an example here, if one knows the detailed application, one can develop tests of the integrated system using relevant parameter values. Without such knowledge, full-system testing may be extremely difficult because of the number of possible combinations.

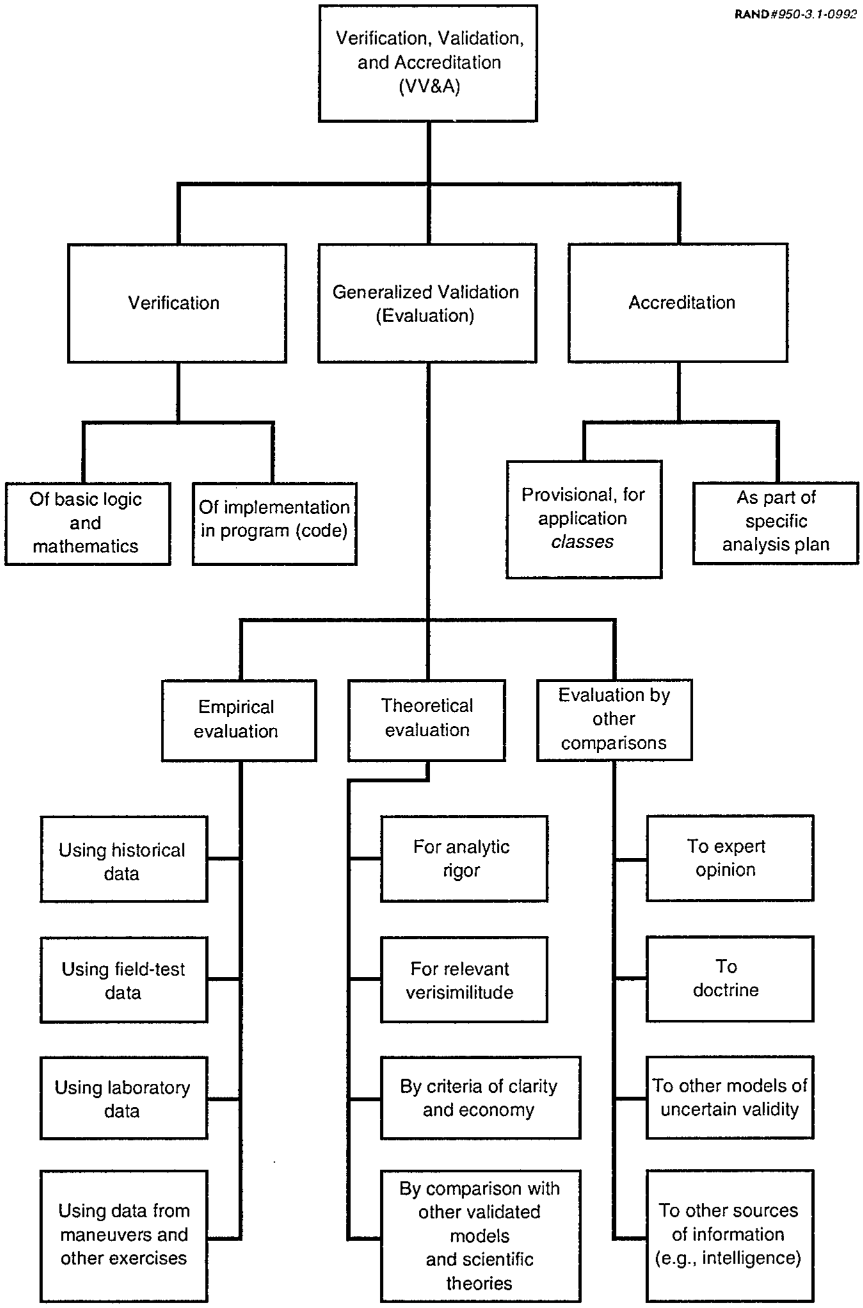

RAND#950-3.1-0992
Verification, Validation, and Accreditation (VV&A)
Verification
Generalized Validation (Evaluation)
Accreditation
Of basic logic and mathematics
Of implementation in program (code)
Provisional, for application classes
As part of specific analysis plan
Empirical evaluation
Theoretical evaluation
Evaluation by other comparisons
Using historical data
Using field-test data
Using laboratory data
Using data from maneuvers and other exercises
For analytic rigor
For relevant verisimilitude
By criteria of clarity and economy
By comparison with other validated models and scientific theories
To expert opinion
To doctrine
To other models of uncertain validity
To other sources of information (e.g., intelligence)

Figure 3.1—A Taxonomic View of VV&A

traditional methods include (a) walking through the design and code, (b) studying flow diagrams, (c) checking algorithms, and (d) using CASE tools. Significantly, modern software methods coupled with the development of expert systems to assist verification can greatly improve the quality of models and the efficiency of the verification process (for example, by detecting errors when they are introduced). Many of the methods seem mundane when described, and may seem burdensome to those who must do the typing of code, but they are exceptionally powerful and have not yet been fully exploited. Examples with which I am personally familiar include:[2]

- Strong typing in computer languages, which detects a wide variety of typographical errors and ambiguities such as having different names for the same variable or different variables with the same name.
- Range constraints on variable values, which are entered (as data) at the time variables are declared and which allow the executing program to become aware of likely errors (as evidenced by variables taking on values outside the prescribed ranges) and to print error messages.
- Automatic testing for logical completeness in decision tables and equivalent sets of If-Then-Else loops.
- Well-structured "explanation logs" at alternative levels of detail, which allow a reviewer quickly to scan not only final results but values of intermediate variables and the logical paths being taken in the simulation.
- Use of object-oriented design methods, which, when physically natural, provide improved modularity and better organized data structures that simplify verification.

These techniques[3] can be especially useful for verification of implementation in code, but can also be useful in highlighting spurious logic (e.g., in explanation logs).

[2] See Zühtü and Ören (1986), Sargent (1986), and Ören (1986) for discussion of ambitious ideas going beyond the examples given here.

[3] Most of these techniques require an "active data dictionary," which is a data base of information on the model's data (e.g., type, format, acceptable values, and meaning). Except for "meaning," the information can be used automatically to check source code and data values.

VALIDATION METHODS

Validation as a Holistic Process

Most experienced modelers and analysts consider validation to be a holistic evaluative process that includes many different kinds of testing. Some of this may be classic empirical testing of the sort often associated with the scientific method. In practice, however, it is only rarely possible in policy analysis to conduct the controlled experiments necessary for such rigorous testing of the model as a whole. Where such experiments are feasible, they should be greatly valued, but we cannot conduct controlled wars or even perfectly controlled battles (nor can we conduct perfectly controlled social experiments on matters such as health care options). We must settle for something a good deal less than idealized scientific rigor.[4] Nonetheless, there is ample opportunity for empirical work. As suggested by the empirical-evaluation column of Figure 3.1, some aspects of models can be tested or informed by comparisons with historical data, field-test data, or data from operational maneuvers and other exercises. These data are not usually as well controlled or as directly relevant as one might like, but they are very useful nonetheless.

Looking to the central column of Figure 3.1, other less empirical methods should be key players in generalized validation. The first is theoretical analysis (e.g., working through the substantive logic, checking relevant verisimilitude, considering the reasonableness of assumptions, applying criteria such as requiring falsifiability[5] and the use of Ockham's razor, and comparing assumptions and implications of the model with well-established theories from physical science, engineering, and military science[6]). Theoretical analysis, then,

[4] Hodges and Dewar (1992) argue that failure to appreciate this reality has been a fundamental source of difficulty in the continuing discussions about validating military models. They argue that the word "validation" should be reserved for predictive models that can be rigorously tested, and that other types of model evaluation should be developed as a function of how the models are to be used (e.g., as bookkeeping devices in a human war game, as decision aids, and as devices to stimulate hypotheses).

[5] It is not uncommon for "theories" to be expressed in ways that make it impossible to disprove them. Good science, by contrast, insists that theories be falsifiable. Indeed, scientists go to considerable lengths to define experiments that stress their theories as much as possible.

[6] As an example of where military science might enter, consider the many theater-level models over the years in which air forces for close air support and battlefield interdiction have not been concentrated in time and space, thereby diluting their potential effect on the other side's ground-force maneuver and ignoring the importance of concentration and coordination to military art generally and to survival and effectiveness of those air forces specifically. As another example, consider the common failure to represent adequately the suppressive effects of artillery. Some models, of course,

goes well beyond what is suggested by the phrase "logical validation," which sometimes appears in discussion of VV&A (e.g., Williams and Sikora, 1991). Theoretical analysis often exploits special cases in which it is possible to compare the model in question with exact calculations based on rigorous or otherwise well-established theories.[7] Sargent (1986, 1987) lists some of the various methods that can be used in this connection.

Looking to the rightmost column of Figure 3.1, one can make a variety of other comparisons to evaluate a model. These include comparisons with expert opinion, doctrine, and so on. Finally, Figure 3.2 emphasizes that these evaluations all feed into an overall evaluation holistically. There is no natural order or ranking of evaluation methods, despite efforts to create one (for example, as discussed ambivalently in Williams and Sikora, 1991, although subsequent MORS works has dropped the effort to impose an order). This is not entirely trivial, since false ideals cause trouble and the ideal of believing, for example, that data from maneuvers are the "best" and "most important" data to be used in validating a model will typically be wrong. Basically, model development and evaluation involves using many sources of information and tying it together however one can. It is not so orderly as some would have it.[8]

handle both of these issues relatively well, but many military models have grossly misrepresented the phenomena, often without justifying their simplifications through auxiliary calculations. Detecting such problems is arguably a matter of "science," not logic or analytic rigor.

[7] It is striking to note that theoretical evaluation is commonly (almost always) omitted from discussion of validation methods. It is most assuredly not the same as "logical verification" or "logical testing." My own sense is that the omission is another symptom of military modeling suffering from not being part of a military science. It has perhaps been overly influenced by mathematicians and programmers, without the emphasis on phenomenology that scientists are supposed to bring to the table (but scientists can also be beguiled by simplistic but elegant mathematics). An important role for military officers, including retired general officers serving as consultants, is to insist that modelers pay more attention to the *real* phenomena. They must demand more military science if the models are to be faithful to their needs.

[8] In MORS work the distinction has been drawn between "output validation" and "structural validation." One can map the activities of Figure 3.2 into these terms, but not neatly. Theoretical evaluation includes both structural validation and testing behavior (outputs) in various special cases that are understood with prior theories or for which there exist solid empirical data. Empirical evaluation in Figure 3.2 relates to output validation in MORS terms. "Other comparisons" in Figure 3.2 involve both structural and output validation. For example, comparisons to expert opinion and doctrine can look at both assumptions and output.

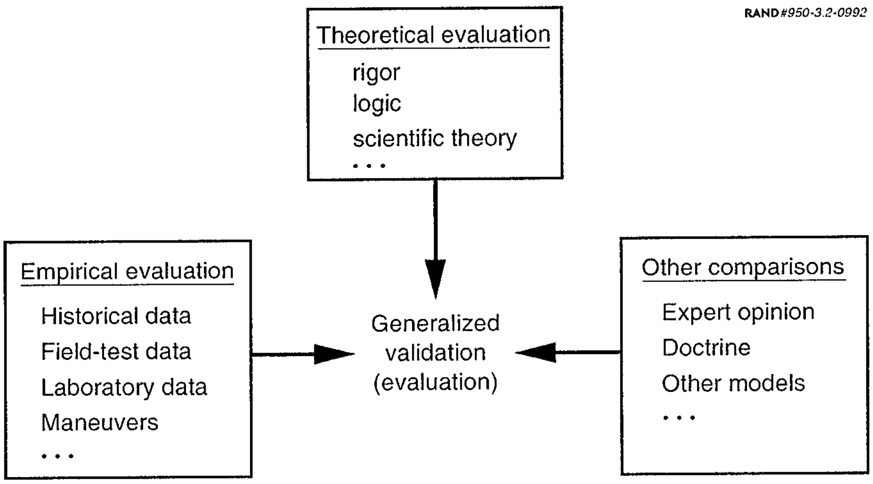


Figure 3.2—Validation as a Holistic Process

A Perspective on Validation

It is sometimes useful to think about validation as an informal application of Bayesian reasoning under circumstances in which we can only estimate the probabilities. Our objective is to develop representations that are good enough "to bet on"; but we will seldom have a sure bet and we therefore want to have a sense of the odds for each of a number of very different kinds of wagers.[9] This validation process is unquestionably subjective, but not capriciously so. We consciously seek information that could falsify or reinforce our judgments and we attempt to face up to that information when we obtain it. When all is said and done, however, we must *do something.* That is, we must conduct the best analysis possible given the information, time, and resources available to us. Ultimately, *validation (and accreditation as well) is concerned with establishing that we are indeed doing the best we can*—or, at least, something that is "good enough." It cannot be separated completely from context.[10]

[9] This view treats validation as a matter of degree. Hodges and Dewar (1992) take a different approach.

[10] As one reviewer of this report noted, "doing something" sometimes should mean doing the best analysis possible even though that means *not* using a computer model that sponsors and users of the computer model are expecting will be used. This may be logically obvious, but it can be a problem in practice because there are instances in

Issues of Breadth and Depth in Model Validation

A model's validity is one thing; the extent to which it has been validated is another (i.e., a good model may not yet be known to be good). A common question for those overseeing the development and use of models is "How much validation is enough?" Another is "How do we start?" Figure 3.1 provides a checklist of methods, but pursuing most of them could become lifetime careers when dealing with complex models. It is therefore useful to make some further distinctions, which also have the effect of suggesting where to start.

As with most human endeavors, the value of validation activity is described by a curve of marginal returns—a curve that rises steeply and then begins to level off and move slowly toward an asymptote (which may correspond to considerable, and yet incomplete, confidence). For a variety of reasons, some of which could probably be explained theoretically, it seems to be the case that even a little validation can go a long way. It is for this reason that "face validity assessments" are so important in practice. These can be attempted in each and every validation-related box of Figure 3.1. Some examples will probably convey the ideas. Once again I use the technique of plausible statements that might be made in characterizing a model's validity:

> *Using historical data.* The model is absurd. It took me all of 30 seconds to discover from the output graphics that it has field armies moving at an average speed of 150 km/day over the course of a successful ten-day campaign. Probably, some nitwit physicist built the post-breakthrough movement algorithms after thinking about how fast tanks can drive. Historically, opposed movement has been more like 20 km/day, although there have been special cases.[11]
>
> *Using field-test and exercise data.* The model is exceedingly optimistic about the effectiveness of TOW missiles (kills per shot and shots per battery per battle), probably because of using test-range data uncritically. Results from the National Test Range and Desert Storm give a very different picture.
>
> *Using simulator data (a kind of laboratory data).* The model for pilot acquisition rates in finding mobile targets is in fact more reliable than what the pilots are telling us anecdotally based on normal training practice. There have been some experiments in simulators that demonstrate pilots are much more conservative about declaring a target detec-

which reference to a well-known computer model is thought somehow to confer a sense of validity, legitimacy, or acceptability.

[11] MacQuie (1987) compiles an interesting array of historical data to be used in tests of face validity. The Army's Concepts Analysis Agency has a continuing effort to exploit historical data (see Helmbold, 1990, for references).

> tion when they are concerned about friendly forces being in the region or about hitting civilian targets. In terms of the required signal-to-noise ratio, the difference is . . .
>
> *Testing for analytic and scientific rigor.* I quit reading the documentation as soon as I discovered that the detection model assumes a uniform background over areas as big as middle-eastern countries. We know that the ability to track a target (not just detect it once) depends on being able to maintain a reasonable signal-to-noise ratio, and that background varies substantially over distances of tens of meters, even in the desert. I also note that the model ignores the effects of cueing and prior knowledge by using independent probabilities. We need a better acquisition model.
>
> *Looking for relevant verisimilitude.* The model treats logistics quite crudely, at the level of tons per day of consumption, tons on hand (by sector), etc. However, it looks about right in aggregate: Divisions in intense combat use about x tons per day, but intensity seems to drop fairly quickly, which is reasonable. The real problem is that there is no mechanism in the model for one side to affect the other side's supply capability. The model is structurally unsound in that respect. It doesn't even model support units and allow attacks on their trucks.
>
> *Evaluation for economy.* The model may or may not be accurate if one knows all the input variables precisely, but it's going to be impossible to use well for systems analysis in realistic cases where we often don't know those values. The model has so many tuning parameters it could fit anything after the fact, but I don't think it's worth much for our purposes.
>
> *Comparisons with familiar models.* Well, it's a different model, of course, and there are scores of parameters that I didn't try to review in detail, but the model at least behaves reasonably in the sense that it gives the same picture of what would happen in the several baseline cases of the . . . study as came out of the full-up war game at CINC headquarters.

All of these examples could have been the result of fairly casual checks of face validity by different experts. None involved detailed testing. In my experience, tests of face validity, in many dimensions, is extremely valuable in uncovering the most serious errors. It is a prerequisite, however, that the model be well documented and that it be easy for experts to view its behavior (for example, through interactive postprocessing graphics rather than fixed hard-copy outputs).

Methods of face-validity testing depend heavily on such things as the following:[12]

[12] Even more fundamental is the need for professional model development practices emphasizing module-by-module testing by developers as a routine part of everyday

- Having a good set of baseline cases (standard scenarios) with which the reviewers are familiar;
- Displays of *aggregated* behavior (e.g., total divisions deployed in theater vs time or average divisional loss rates when in combat vs time);
- Highly organized and comprehensible overviews of model approach, assumptions, and parameter values (more generally, good documentation is essential; see Appendix B for more discussion of documentation);
- The ability to respond quickly to spot-check requests (e.g., "What did you assume for the value of . . . ?" and "What does the plot of . . . vs time look like?" and "Show me, in code, the algorithm (or rules) you used for . . . ")
- The ability to do additional spot-checks on demand (e.g., "Let's see what happens when you assume the B-1B's ECM doesn't work").

The dangers of depending only on face validity are obvious, but they can be mitigated if the effort to do face-validity checks is broad enough, includes opportunities for spot-checking in depth, is accomplished with reviewers having a range of backgrounds, and mixes review of "inputs" (model structure, assumptions, etc.) and "behavior." One reason such testing is so valuable is that poorly done models often fail immediately, whereas well done models are the result of serious and professional efforts in which testing and validity-related discussions are an everyday way of life for developers. Given such efforts, intensive review sessions can cover a great deal of ground quickly because the developers are "on top of the problem" and have organized information well.

Detailed validation efforts must depend primarily on module-by-module testing during development and on special meetings to examine critical modules in depth. It is seldom possible with large military models to do anything like comprehensive testing or evaluation of complete multimodule systems.[13]

work. If more sloppy methods have been followed, face-validity efforts are likely either to fail or be quite misleading.

[13] An important point is that much more extensive testing *would* be possible if it were budgeted. It is unusual, however, for military simulation projects to set aside, for example, 20 percent of the overall project funds for independent and comprehensive VV&A. In some instances, such testing would be well worth the investment. In many other cases, however, some imperfections are quite tolerable.

Special Issues with Knowledge-Based Models

Knowledge-based models such as rule-based or algorithmic and rule-based decision models representing, for example, military commanders or operators of air defense systems raise special issues because in most cases they cannot in principle be validated in the sense of being favorably compared with "the real system." Instead, they must be evaluated on grounds such as whether they faithfully represent the knowledge of relevant experts, whether they are logical, internally consistent, and consistent with various physical and logical constraints, and so on.[14] They can in some cases be falsified by real-world experience in which other variables proved to be critical, but ambitions must be limited. Further, there is a wealth of information to the effect that experts often give misleading testimony about what they would do in various circumstances and about the way in which they reason—not because they *intend* to mislead, but because they have only a limited understanding of their own cognition. For example, when being interviewed experts might describe a highly rational process of making decisions, but in the heat of actual operations—with uncertainties, fatigue, and time pressures all being factors—their behavior might reduce to the simplest of patterns, some of them "irrational" from the viewpoint of a decision theorist. To make things worse, most experts have never encountered many of the situations for which we may be asking them to predict behavior. Thus, they are not *really* experts in the same sense that an experienced internist is an expert on childhood diseases.

It follows from this that efforts to validate knowledge-based models, notably behavioral models of various types, including decision models, must depend much more heavily than one might like on combinations of theory, logic, and spotty expressions of expert opinion.[15] It is essential that efforts to build such models be highly organized and that appropriate testing methods be developed. This is an understudied field, but some relevant methods that have been applied in a number

[14] Some concrete examples here come from a recent evaluation by the Center for Naval Analyses (CNA) of a command and control model. The review asked: (a) Have all the decision nodes been identified? (b) For each node, has a variable been defined for each factor that could affect decisions at that node? and (c) For every possible state of each variable at each node, has a rule been developed (e.g., an If/Then statement) and does the rule reflect the judgment of experts?

[15] My own experience with knowledge-based models has emphasized theory and logic, with experts being used mostly for spot-checking (e.g., Davis, Bankes, and Kahan, 1986). The textbook concept of using "knowledge engineers" to extract knowledge from experts often does not apply or is less efficient and organized than having a subject-area analyst build a model and then iterate it by talking with experts. For a discussion of the knowledge-engineering approach, see Waterman (1986).

of domains are described in Veit, Callero, and Rose (1984) and Veit (forthcoming). These involve developing rigorous factorial designs for comparing model behaviors with behaviors of relevant experts, preferably in circumstances approaching those that would be encountered in the real world, but perhaps in war games as a next-best choice. Another valuable empirical approach is to observe experts performing in field exercises. This can usefully supplement interview data and theoretical analysis by bringing in, to some extent at least, aspects of behavior under stress and the fog of war.

METHODS OF ACCREDITATION

There are various organizational approaches to accreditation, but this subject is best discussed in the next section.

4. A DYNAMIC VIEW OF VV&A

Figure 4.1 shows a dynamic view of VV&A that emphasizes evaluation and accreditation of a model in the context of a specific study.[1] The importance of context is emphasized by putting the analytic plan in the center. It is here one starts—knowing, of course, the purposes of the analysis. Provisional accreditation for a *class* of applications could emerge from a similar chart, but I will not deal with that further in this study.

When evaluating a model for a specific application, chances are that the model is an adaptation of a previous model that has been subjected to some degree of VV&A or subjected previously to considerable "general" VV&A without the benefit of study-specific information.[2] Thus, the new round of VV&A shown in Figure 4.1 draws on previous information (see the arrows coming in from the top left). Most important, however, it depends heavily on the study-specific requirements and test cases. In practice, relatively complex combat models (or most other models used in policy analysis) are never *fully* tested and unconditionally accredited. Testing can still be extensive and sophisticated for the purposes of evaluating the model and its data in the context of a specific analysis. That testing is the basis for study-specific accreditation, but it also adds to the base of VV&A information that will be used in the next iteration for a new application (see the outward arrows on the bottom left and center right).

One feature of Figure 4.1 (bottom right) is especially important and unusual. This is its reference to constraints and guidance as outputs of the accreditation process. Since the most stringent review of an analytic organization's work usually occurs within the organization

[1] This discussion envisions a model being used for an analysis study. However, analogous diagrams could readily be constructed for such other applications as training, education, and operational decision aids. Some readers may wish to do so.

[2] There is an issue of balance and complementarity here. Some discussions of VV&A convey the impression that models can be adequately evaluated once and for all, when in reality model appropriateness must be judged in the context of an application. However, studies often face time pressures and modest resources, which means that they cannot take on the full burden of evaluating models from scratch and depend on there having been a considerable degree of prior VV&A. While Figure 4.1 deliberately focuses on VV&A for an application, both that and the broader VV&A are increasingly considered essential (e.g., U.S. Army, 1992). Personally, I would argue that generic V&V is essential, and generic accreditation is potentially useful (and potentially troublesome), depending on organizational sophistication, integrity, and efficiency.

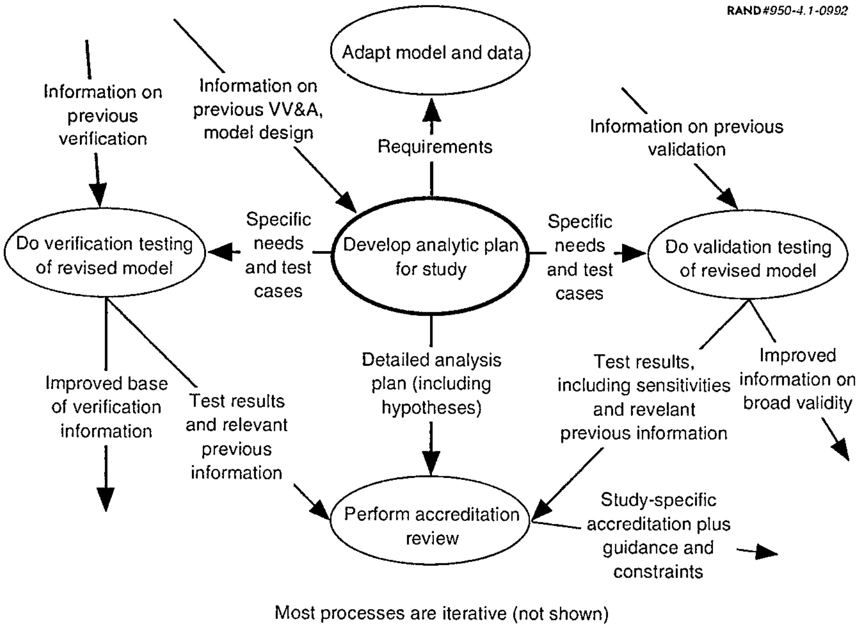

RAND#950-4.1-0992
Adapt model and data
Information on previous verification
Information on previous VV&A, model design
Requirements
Information on previous validation
Do verification testing of revised model
Specific needs and test cases
Develop analytic plan for study
Specific needs and test cases
Do validation testing of revised model
Detailed analysis plan (including hypotheses)
Test results, including sensitivities and revelant previous information
Improved information on broad validity
Improved base of verification information
Test results and relevant previous information
Perform accreditation review
Study-specific accreditation plus guidance and constraints
Most processes are iterative (not shown)

Figure 4.1—VV&A as a Continuing Process Sensitive to Context
(process starts at the center)

itself, one may think of "accreditation" as being the result of management reviews of the sort that should occur early in a project's life, before the project's work is reported, and, if possible, at least once in between. The result of such a review might take the following form (think of this as the summary conclusions of the relevant manager, who need not be a government official):

> On balance, our conclusions are:
>
> 1. The analytic plan appears to be sound.
>
> 2. The model and data base for carrying out the plan appear to be sound.
>
> 3. Consistent with the improved plan, however, no conclusions should be drawn regarding . . . , because the analysis cannot support them. Further, in drawing conclusions on . . . , it is essential that they reflect parametric variations on the following key variables over the ranges discussed in the review. Recipients of the analysis must understand the considerable uncertainty associated with. . . .

> 4. Further, recipients of the analysis must be reminded of the following basic assumptions of the approach, which appear reasonable, but which also establish limitations on its significance: . . . .

In this depiction *there is no all-or-nothing blessing of the model—even for a specific study*. Instead, the accreditation is conditional upon the analytic plan itself, which includes the proposed logic to establish conclusions. Further, the accreditation process often results in changes of the analytic plan itself (and changes in the model leading to another round of verification). This iteration is merely implicit in Figure 4.1.

In concluding that a model could reasonably be used for the purpose at hand, the accrediting authority might be drawing on highly study-specific information and pondering in some detail precisely what function the model itself is serving (see Hodges and Dewar, 1992, for a list of such functions and related discussion).

One can imagine judgments such as the following being made as part of the accreditation decision and explanation:

> The model is suitable here (e.g., in a war game being used for higher-level education and training). Realistically, it is being used primarily as an organizing device, as a kind of bookkeeping mechanism. The results of the analysis depend most sensitively on the human command-control decisions, including operational strategies. The model's treatment of attrition is fairly crude, but as you have shown with your sensitivity analyses, the attrition model is not the limiting factor.
>
> The model is quite suitable here, despite its exceptionally simple treatment of close combat. The results depend primarily on the air-to-ground effectiveness of U.S. air forces, given air supremacy, and the time required for us to achieve that supremacy. You have a rather detailed and credible treatment of both air-to-ground effectiveness as a function of circumstance and of the suppression of air defenses (SEAD).[3]
>
> You must be kidding. The model can't possibly be used to infer conclusions about the proper mix of tank and artillery units, because it bases ground combat attrition on some aggregation expressions that treat the

[3] In a similar spirit, a colleague and I conducted a study of possible post-crisis defense requirements a few months before the allied offensive against Saddam Hussein, in which we used an extremely simple spreadsheet model using Lanchester equations and aggregated force strengths for ground combat. Thę reason for doing so was that we observed that results of more sophisticated and complex war gaming analysis were driven by a few factors (e.g., air-to-ground effectiveness) and that these factors were being obscured by the original level of detail (see Shlapak and Davis, 1991). For other purposes, however (e.g., evaluating *offensive* capabilities), the simple model would have been ludicrously inappropriate.

> multiple launch rocket system as merely one contributor to an overall firepower. Chances are the model will conclude something like "all we need to do is buy MLRS batteries and disband the rest of the army." That would be fine if battle were just a matter of firepower.
>
> Yes, I know that you think you have a highly sophisticated model of ground combat, but it is not adequate for this study. As it stands, ground forces are unintimidated by air forces, and can maneuver just as quickly with or without enemy air forces attacking them, except to the extent that air forces can destroy whole units. I don't believe this for a moment. Air forces can disrupt and delay, and thereby greatly affect maneuver and tempo generally. Go back to the drawing boards—and read some history on the Battle of the Bulge, especially the part after the weather cleared.
>
> Your model seems fine so far as it goes, covering attrition and movement processes, but it treats operational strategy as input data and doesn't allow adaptation. That leaves out the most important part of force employment. Good forces and bad strategy lead to bad results (see, e.g., Davis and Hillestad, 1992).

An important point to be made here is that the same model might be good for some force-composition or force-structure studies and altogether inappropriate for others. Thus, attempting to accredit a model for whole classes of studies can readily lead to bad decisions. It would therefore seem appropriate to introduce and use the concept of *provisional accreditation*, suggested to me by Clayton Thomas: "This model (and its data base) is a reasonable candidate for use in this kind of study. Go ahead and flesh out the analysis plan and let's then see whether the plan makes sense and the model will indeed be adequate." This emphasizes yet again that it is the analysis, study, or other application that should actually be "accredited."

5. ESTABLISHING A VV&A REGIME WITHIN AN ORGANIZATION

PREFACING COMMENTS

In thinking about VV&A and about how to improve its practice in organizations, it is important to recognize that VV&A should not be seen as a separate and segmentable enterprise—i.e., an additional duty or task—but rather as an inherent part of the analytic process from the time of initial design to the time of particular applications. Validation is *central* to the scientific process that good analysis seeks to emulate. I raise these matters here because VV&A is not always viewed in this way. Indeed, many considerations undercut attempts to make analysis "scientific." For example, models are often tools of advocacy; further, data bases are often tightly held for both security reasons and information-is-power reasons. As a result, organizations face significant *disincentives* to evaluate their models and data as harshly as they might if they were physical scientists attempting to unravel the secrets of the universe. It is therefore a significant challenge for analytic organizations to rise above these problems and instill and maintain a sense of professionalism and "scientific method." This is a continuing challenge, not one that can be addressed once and for all (see also Hughes, 1989, pp. 10 ff). With this background, then, let us examine how an organization might take on the challenge.[1]

CONSIDERATIONS

Establishing a VV&A regime must first be recognized as involving all of the standard challenges associated with organizational change and learning. Simple decrees have very limited and short-term value. Instead, one must think in terms of such matters as:

- Creating and communicating a *vision* of professionalism that treats VV&A as inherent to good work and something to be done

[1] Although not discussed in this study, a major issue is how the DoD can create positive incentives for VV&A. Currently, most of the "incentives" under discussion are in the nature of requirements and threats. The most obvious incentive, however, is money: By budgeting appropriately for serious VV&A, the DoD would quickly find itself receiving first-rate proposals for high-quality testing. The second principal incentive I see is the fostering of an invigorated military science as discussed in Davis and Blumenthal (1991).

continuously rather than merely in occasional painful and unrewarding crash efforts.

- Developing associated policies and procedures, and assuring that there are early examples for everyone to see of how these will be implemented in practice and what will be accomplished.
- Bringing members of the organization into the problem so that they participate in developing aspects of the general policies and many of the procedural details—thereby assuring proper tailoring to the organization's particular culture.
- Establishing the uncomfortable principle of *independent* review, at least for critical features of the work, even though the tendency within organizations is usually to assume that internal review is quite adequate and that the call for independent review is insulting and a potential waste of time.[2]
- In all of this, having both long- and short-term views and plans, with short-term efforts being designed in part to illustrate what is intended on a continuing basis for the long term.
- By distinguishing short- and long-term plans, assuaging fears about unreasonable new demands being added immediately to project burdens.
- Assuring that those contributing to the changes are properly recognized and rewarded.

Many aspects of this challenge can be helped by having concrete examples to use as case histories that everyone reads. An important part of the continuing MORS effort on VV&A is to develop and, if possible, to publish such histories.

[2] There is a strongly held view in the larger software community that good VV&A is necessarily *independent* VV&A. Indeed, it is not uncommon to have separate organizations charged with development and VV&A. The motivation here is recognizing that developers often have profound conflicts of interest that undercut VV&A. The pressures include deadlines, cost, the desire to include new and more sophisticated submodels, and the antipathy of workers for the drudgery of extensive testing. An independent tester paid specifically to certify software has, by contrast, other incentives. At the same time, there is substantial evidence demonstrating that "independent testing" cannot usually be conducted in isolation: It is essential for the testers to interact with both developers and users. Developing appropriate working relationships that balance independence of judgment with cooperation and exchange of information is therefore important.

USING THE FRAMEWORK

Against this general backdrop of challenges, I suggest using the material of this study as follows:

- Use the definitions and related discussion to communicate the fundamental issues of VV&A.
- Use the taxonomy of VV&A methods (Figure 3.1) to broaden perspectives, break down biases, and help establish short-term and long-term plans. In the long-term plan, for example, one might want to use *many* of the validation techniques mentioned, but that would require scheduling and finding support for tasks, or even whole projects, for work that would not ordinarily be done at all (e.g., comparisons with experiences in field maneuvers or large-scale exercises). Thus, the taxonomy should be used primarily as a *checklist*.
- Use the dynamic view of VV&A (Figure 4.1) to frame the issues in a realistic, technically solid, and "nonpolitical" way. Use it also to develop detailed work schedules for projects—setting aside adequate time for iterative reviews and follow-up model adaptation and testing. Use this view of the problem to highlight the substantive rôle of accreditation (as distinct from the more political role emphasized by cynics) and its intellectual relationship to traditional guidelines on how to run analysis projects—guidelines that apply also in many ways to applications such as support of exercises and development of decision aids.
- When identifying VV&A requirements for a particular analysis, explicitly consider the costs of fulfilling those requirements. Then, either assure that the requirements can be met by making available the necessary resources and calendar time or adjust the analyst plan (or claims made about the analysis when concluded).[3]
- Take seriously the discussion of how special measures need to be adopted in evaluating knowledge-based models and other models for which hard data are lacking. Use the examples provided here and develop important distinctions for the problems at hand.
- Use Figures 3.1, 4.1, and related discussion to explain to sponsors how VV&A plans are consistent with a comprehensive view of the subject, drawing also on other published materials such as Sargent (1987) and methods used by Martin Marrietta (1990). As part of

[3] The issue of budgeting for VV&A is fundamental, and the failure to appreciate this probably underlies many of the VV&A problems in the military modeling community.

this, focus sponsors and accrediting authorities (usually the same individuals) on the view of accreditation that encourages them to provide intellectual guidance, not merely a "yes" or "no" decision. And, as part of this, emphasize the need for VV&A activities to be adequately supported and scheduled realistically over time.

Finally, let me mention again that the examples in this study emphasize applications in which models are used for analysis. Many readers will wish to develop analogous examples for their own applications, which may relate to training, education, operational decision aids or other matters. While the basic framework should hold up, the detailed criteria for judging models depend very much on application.[4]

[4] As one example, consider that program planners often think in terms of aggregations that are of little or no value to officers participating in operational exercises. As a result, they need different models. Ideally, the models will be consistent, but that is not always easy (Davis and Huber, 1992).

Appendix A
ON SEPARATING CONCEPTUAL MODELING AND PROGRAMMING

In a classical ideal with which I long had sympathy, the design and review of models (sometimes called conceptual models) precedes programming.[1] One develops the conceptual picture and lays out the theory and algorithms formally, thereby creating machine- and language-independent specifications (see, e.g., Figure A.1 from Sargent's work, which remains useful even if my arguments here are accepted). Implementation as a program then proceeds, but its details depend on hardware, software, local practices, and other factors.[2] In this ideal, substantive discussion should focus on the model, not the program. This ideal has much to recommend it, because enormous confusion is caused by having problem formulation shaped and described in terms peculiar to particular languages or computer systems.

In practice, however, the ideal breaks down for both good and bad reasons. The principal bad reason is that many organizations lack the discipline to enforce serious design before allowing programmers to write code. The results are predictable: incomprehensible models that are merely implicit in long and complex computer code.

The good reasons have to do with technology and the changing ways in which workers, even workers with a theoretical bent, go about their efforts. It is becoming increasingly possible and attractive to work largely at the computer rather than with pencil and paper—even for constructing top-down conceptual designs. Second, some of the computer tools for doing so blur the distinction between design and programming, because when one creates the initial design elements (e.g., variable names, data structures such as objects, functions, and diagrams), the results automatically generate corresponding program elements (see Appendix B). Third, with some high-level

[1] See, for example, Zeigler (1984), Sargent (1986, 1987), and Martin Marrietta (1990).

[2] As discussed by Julian Palmore of the University of Illinois in an address to the 60th MORS conference in Monterey, California, in June, 1992, even details of computer arithmetic can be very important in simulation. Failure to pay attention to such details can produce substantial "structural variance" as manifested, for example, by peculiar sensitivity results and major changes in results if one shifts from one computer to another. See also Palmore (1992).

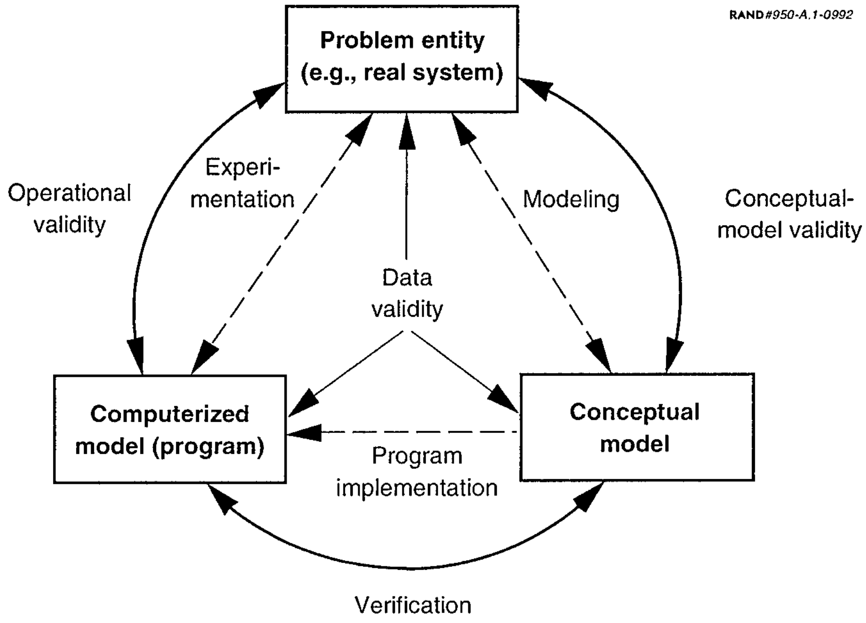


Figure A.1—An Idealized Separation of System, Model, and Program

languages, it is as easy for reviewers to understand and comment upon algorithms expressed as computer code (or related diagrams) as it is for them to do so in a paper-and-pencil mode.[3] Fourth, advanced tools such as *Mathematica*™ now make it possible to solve equations symbolically on line, which enhances the design process. And, last, statements of the conceptual model often underspecify the problem, resulting in programmers filling in and thereby having much more of a role in defining the "real" model than was intended. In some respects, it is only realistic to force model designers to address explicitly what they might otherwise tend to assume are mere implementation issues (e.g., time steps, control flow in procedural problem-solving ap-

[3] Separate documentation is still needed for gaining a top-down overview of the model and program. Further, it is virtually essential when the program itself is large. However, the documentation may be out of date or may contain errors that do not exist in the code (and, of course, the code may contain errors not in the documentation). My own view is that future reviews of models should ideally combine reading of documentation for top-down structure and having that documentation, which may also be on line, "point to" critical portions of code that can be examined directly. That will be increasingly feasible with high-level computer languages and environments (see Appendix B).

proaches, and whether to organize around data structures or processes).

A related issue here is that of prototyping. In the last decade workers have come to appreciate the efficiency of rapid prototyping as a mechanism for helping designers understand the problem for which they are tasked to build models. In practice, it is common for even first-rate modelers and analysts to misunderstand major elements of the problem until they have actually built something and worked with it. While preliminary design is necessary, it is seldom sufficient and those with modern software tools tend strongly to recommend highly iterative development that exploits prototyping and the discovery process as an inherent part of high-quality work, not something to be apologized for.

While I continue to recommend separating model design from design of detailed implementation, and while I still believe it is desirable for many aspects of a model to be reviewed away from the computer context, which tends still to encourage a linear line-by-line view and inelegant solution techniques, the original ideal is now, in my view, obsolete. It is a major challenge for developers to create new operating procedures that will maximize benefits of computer environments while maintaining an appropriate separation of model and implementation detail.

Appendix B

DOCUMENTATION, HIGH-LEVEL COMPUTER LANGUAGES, AND MODERN MODELING AND ANALYSIS ENVIRONMENTS

DOCUMENTATION

A prerequisite for VV&A is documentation, but many DoD combat models are inadequately documented. To improve this situation, it is important to know what constitutes good documentation. The DMSO's Applications and Methodology Working Group discussed this at some length in 1991, drawing heavily on the experience of participants, many of whom had actually developed large models or had evaluated them in detail. It agreed that the following guidelines are especially important:

- Distinguish the model from the program (i.e., describe the conceptual model in terms that are language independent and focused on the underlying concepts and relationships);
- When appropriate, describe the model in object-oriented terms, even if the implementing program is not object oriented;[1]
- Require high-level designs describing motivation, rationale, and basic assumptions, plus:
  - Hierarchical top-down structures (where hierarchies apply) and data-flow diagrams to show how inputs get transformed into outputs;
  - Meanings of variables (input to data dictionaries);
  - Logical or algorithmic detail on selected key modules;
  - Structured and commented on source code, even though this cannot replace documentation, especially higher-level documentation;
  - Program and interface documentation and illustrative-scenario "walkthroughs."

[1] One can design a model in terms of objects, attributes, processes, and the like whether or not the programming language has the paraphernalia of objects, messages, methods, and so on.

Distinguishing the model from the program is important to sharpen and communicate concepts, even if the arguments of Appendix A are accepted. Programmers often talk about pointers, memory, stacks, arrays, and other constructs having nothing to do with military phenomenology. Documentation and reviews of model content should instead focus on phenomenology.

One important element of good documentation is often overlooked: including the procedures and results of any previous VV&A efforts conducted during development or applications. This can be exceptionally useful.[2]

There are limits to how much documentation can be squeezed out of money-limited projects. The most important documentation consists of "high-level designs," which are top-down in character with an emphasis on structure. These should also define key variables, provide appropriate diagrams showing, for example, information flow and control flow, and provide logical or algorithmic detail on key submodels. It is less important, and may even be inappropriate, to document details of much of what constitutes a complex combat model, since those details are often bookkeeping methods best understood at the level of the code itself. The code, however, should be well structured and commented on. Another major element of documentation is information on how to use the program and its interfaces. This is often best done by providing a step-by-step discussion of how one runs and analyzes a test case (i.e., a walkthrough of a representative application in a given scenario). Commercial software tools often have excellent "walkthrough" documentation.

Taken together, then, there is need for documentation on the model, the program, and its use. Increasingly, on-line documentation is becoming especially important for procedural information.

Finally, note that documentation methods should be changing, and that should be reflected in work on comprehensive environments.

HIGH-LEVEL LANGUAGES AND ENVIRONMENTS

The phrase "high-level language" is ambiguous, because there are multiple dimensions along which to measure. SIMSCRIPT was one of the first high-level languages designed for simulation. It was high-level in such respects as providing tools making it easy to construct

[2] In naval modeling a special need is discussion of how environment is handled in the model.

simulations. It also had mechanisms to force good programming practices such as writing an overview of the model, using descriptive identifiers, and exploiting class concepts. In more recent times, spreadsheet languages such as Excel may be considered very high level in the sense of having user-friendly interfaces and a myriad of predefined functions. At the same time, spreadsheet programs are usually the antithesis of structured programming, because the approach taken by the novice is to organize by spreadsheet cells and use the equivalent of many GO TO statements producing "spaghetti code." Further, complex spreadsheet programs based on the systematic use of macros are no more intelligible than those of other languages such as BASIC, and arguably less so.

Against this background, RAND has been developing high-level languages that emphasize using relatively natural language for key words and that exploit the cognitive effectiveness of table structures for organizing both information and logic. RAND now has seven years of experience with RAND-ABEL®, which has been used to write hundreds of thousands of lines of code. The applications have ranged from decision models (e.g., those of a simulated theater commander) to combat models (e.g., attrition and movement processes for combat taking place on a network). It has consistently proven possible to have group reviews of major portions of these models by working directly with code, even though many of the participants have not been serious programmers. Errors have been discovered at a glance, and complex logic has been discussed as a group. Most of this has been possible because of the table structures, which should be developed in other languages as well.

In current work, RAND is developing an object-oriented version of RAND-ABEL, called Anabel.[3] This will extend the effort to exploit two-dimensional structures of many kinds (e.g., decision tables, tables of orders, and adjudication tables) and will also include numerous self-documenting features, including the use of hyper media. Our belief is that model documentation will not improve greatly by virtue merely of managers cracking whips. Instead, there is both need and opportunity for technology to help. Similar ideas are being pursued at many levels by a variety of researchers, including some who are contemplating the use of expert systems to help choose and use verifi-

[3] Anabel, the result of ideas by Edward Hall and Norman Shapiro, is being developed as part of a grander scheme for a modeling and analysis environment (see Anderson, Bankes, Davis, Hall, and Shapiro, forthcoming). RAND-ABEL is documented in Davis (1990) and Shapiro, Hall, Anderson, LaCasse, Gillogly, and Weissler (1988).

cation and validation tools (see, for example, Ören, 1986, and Sargent, 1986). In addition, researchers are developing a variety of excellent graphical tools, some of them capable of generating c ode directly. The Systems Dynamics programs Stella® and iThink® are especially notable here. Plans call for a variety of such tools to be used with RAND's Anabel, building on tools recently developed by Larry McDonough and Richard Hillestad. One, called Mapview, allows workers readily to create objects and emplace them on maps. The results of what they do with the graphical interface generate code. Similarly, a tool called the Activity Sequence Editor (ASE) allows workers to develop state-transition diagrams for object-oriented programs, and to have the results of those diagrams generate code. All of this facilitates documentation and VV&A, because many aspects of model design are best seen graphically, and because the tight linkage between diagrams and code avoids the traditional problem of documentation lagging the reality embedded in the code itself. Despite the progress, however, there is a great deal to be done in this general subject area.

A THREAT TO ADVANCEMENTS

Progress in developing and disseminating advanced modeling and analysis methods and tools, including many that would facilitate VV&A, will be adversely affected if the DoD attempts to force all modeling activities into a single structure and language, such as Ada in particular. Such a policy would hinder efforts to exploit the rich selection of commercial products that exist and are emerging. It would also hinder efforts to develop advanced tools, many of which are most readily developed within existing computer environments (e.g., Unix and Macintosh). The motivation for commonality is understandable, and the desire for greater reusability and interoperability of software is laudable, but the requirement for a single language is misplaced. *High degrees of reusability and interoperability can be accomplished with standards that are language independent.* Indeed, that is what makes "open architectures" feasible and important. Ada is a powerful language that can greatly contribute to the management and control of software development in many projects, but it is much less suitable for prototyping, or for models that will continue to change and that deal with highly uncertain phenomena. For such models there is a high premium on, for example, interactiveness, flexibility, clarity, explanation capabilities, and easy connectivity to commercial tools.

BIBLIOGRAPHY

Anderson, Robert H., Steven C. Bankes, Paul K. Davis, H. Edward Hall, and Norman Z. Shapiro (forthcoming), *Toward a Comprehensive RAND Environment for Computer Modeling, Simulation, and Analysis.*

Bonder, Seth (1984), "Summary of a Verification Study of Vector-2 with the Arab-Israeli War," in Reiner Huber (ed.), *Systems Analysis and Modeling in Defense: Developments, Trends, and Issues,* Plenum Press, New York, 1984.

Brewer, Gary, and Martin Shubik (1979), *The War Game: A Critique of Military Problem Solving,* Harvard University Press.

Davis, Paul K. (1990), *An Analyst's Primer for the RAND/ABEL® Programming Language*, RAND, N-3042-NA.

Davis, Paul K., and Donald Blumenthal (1991), *The Base of Sand Problem: A White Paper on the State of Military Combat Modeling*, RAND, N-3148-OSD/DARPA.

Davis, Paul K., and Reiner K. Huber (1992), *Variable-Resolution Modeling: Motivations, Issues, and Principles*, RAND, N-3400-DARPA.

Davis, Paul K., and Richard Hillestad (1992), "Using Simulation in the Education of General Officers," *Proceedings of the Summer Simulation Conference.*

Davis, Paul K., Steven C. Bankes, and James P. Kahan (1986), *A New Methodology for Modeling National Command Level Decisionmaking in War Games and Simulations*, RAND, R-3290-NA.

Defense Modeling and Simulation Office (1991), *Final Report of the Applications and Methodology Working Group,* December 1991.

Department of Defense/Defense Modeling and Simulation Office (1992), *Defense Modeling and Simulation Initiative*, May 1.

Dupuy, Trevor (1987), *Understanding War,* Paragon House, New York.

Elzas, Maurice S., Tuncer Ören, and Bernard Zeigler (eds.) (1986), *Modelling and Simulation Methodology in the Artificial Intelligence Era*, North-Holland, Amsterdam.

Gass, Saul (1983), "Decision-Aiding Models: Validation, Assessment, and Related Issues for Policy Analysis," *Operations Research,* Vol. 21, No. 4, July–August 1983.

Helmbold, Robert L. (1990), *Point Paper: CAA History Activities, 1980–1990*, U.S. Army Concepts Analysis Agency, Bethesda, Maryland.

Hodges, James S., and James A. Dewar (1992), *Is It You or Your Model Talking? A Framework for Model Validation*, RAND, R-4114-AF/A/OSD.

Hughes, Wayne (ed.) (1989), *Military Modeling, Second Edition*, Military Operations Research Society, Alexandria, Virginia.

Klir, G. J. (1989), "Inductive Systems Modeling: An Overview," in Maurice S. Elzas, Tuncer Ören, and Bernard Zeigler (eds.), *Modelling and Simulation Methodology in the Artificial Intelligence Era*, North-Holland, Amsterdam.

Martin Marietta (1990), *Confidence Methodology Guide*, Third Edition, Final, National Test Bed Technical Report NTB-237-022-06-02, August 15. Prepared for Strategic Defense Initiative Organization, Washington, D.C.

McQuie, Robert (1987), *Historical Characteristics of Combat for Wargames (Benchmarks)*, Army Concepts Analysis Agency, CAA-RP-87-2.

Military Operations Research Society (MORS) (1989), *Human Behavior and Performance as Essential Ingredients in Realistic Modeling of Combat—MORIMOC II,* Alexandria, Virginia. (Proceedings of a conference held 22–24 February 1989.)

Miser, Hugh J., and Edward S. Quade (1988), *Handbook of Systems Analysis: Craft Issues and Procedural Choices*, Elsevier Science Publishing Co., New York, New York.

Ören, Tuncer (1986), "Artificial Intelligence in Quality Assurance of Simulation Studies," in Maurice S. Elzas, Tuncer Ören, and Bernard Zeigler (eds.), *Modelling and Simulation Methodology in the Artificial Intelligence Era*, North-Holland, Amsterdam.

Ören, Tuncer (1989), "Bases for Advanced Simulation: Paradigms for the Future," in Maurice S. Elzas, Tuncer Ören, and Bernard Zeigler (eds.), *Modelling and Simulation Methodology in the Artificial Intelligence Era*, North-Holland, Amsterdam.

Pace, Dale K., and Dennis P. Shea (1992), "Validation of Analysis Which Employs Multiple Computer Simulations," *Proceedings of the Summer Simulation Conference*, pp. 144–149, Society for Computer Simulation.

Palmore, Julian (1992), "Analysis and Verification and Validation of Complex Models," *Proceedings of the Summer Simulation Conference*, pp. 139–144, Society for Computer Simulation.

Sargent, Robert (1986), "An Exploration of Possibilities for Expert Aids in Model Validation," in Maurice S. Elzas, Tuncer Ören, and Bernard Zeigler (eds.), *Modelling and Simulation Methodology in the Artificial Intelligence Era*, North-Holland, Amsterdam.

Sargent, Robert E. (1987), "An Overview of Verification and Validation of Simulation Models," *Proceedings of the 1987 Winter Simulation Conference*, Society for Computer Simulation.

Shapiro, Norman Z., H. Edward Hall, Robert H. Anderson, Mark LaCasse, Marrietta S. Gillogly, and Robert Weissler (1988), *The RAND-ABEL® Programming Language: Reference Manual*, RAND, N-2367-1-NA.

Shlapak, David A., and Paul K. Davis (1991), *Possible Postwar Force Requirements for the Persian Gulf: How Little Is Enough?,* RAND, N-3314-CENTCOM/JS.

Shubik, M., and G. D. Brewer, *A Survey of Models, Simulations, and Games*, RAND, RM-7821-1-ARPA.

Stockfisch, John A. (1975), *Models, Data, and War: A Critique of the Study of Conventional Forces*, RAND, R-1526-PR.

Thomas, Clayton (1983), *"Verification Revisited,"* in Wayne Hughes (ed.) (1989), *Military Modeling, Second Edition,* Military Operations Research Society, Alexandria, Virginia. [In the original 1983 edition of Hughes' book, Thomas used "verification" in the sense that is now applied to "validation." He changed his usage for the 1989 edition cited here.]

U.S. Army (1992), Army Regulation 5-11, Army Model and Simulation Management Program, June.

U.S. General Accounting Office (1980), *Models, Data, and War: A Critique of the Foundation for Defense Analyses*, GAO/PAD-80-21, Washington, D.C.

U.S. General Accounting Office (1987), *DoD Simulations: Improved Assessment Procedures Would Increase the Credibility of Results*, GAO/PEMD-88-3, Washington, D.C.

Veit, Clairice (forthcoming), *Methods for Developing Validated Subjective Measures to Incorporate in Simulations,* RAND, R-4209-A/DDR&E.

Veit, Clairice T., Monti Callero, and Barbara J. Rose (1984), *Introduction to the Subjective Transfer Function Approach to Analyzing Systems,* RAND, R-3021-AF.

Waterman, Donald A. (1986), *A Guide to Expert Systems*, Addison-Wesley, Reading, Massachusetts.

Williams, Marion L., and James Sikora (1991), "SIMVAL Minisymposium—A Report," *Phalanx*, Vol. 24, No. 2, June.

Zeigler, Bernard (1976), *Theory of Modelling and Simulation*, John Wiley and Sons, New York; reissued by Krieger Pub. Co., Malabar, Florida, 1985.

Zeigler, Bernard (1984), *Multifacetted Modelling and Discrete Event Simulation*, Academic Press, London and Orlando, Florida.

Zühtü, Aytaç, and Tuncer Ören (1986), "MAGEST: A Model-Based Advisor and Certifier for GEST Programs," in Maurice S. Elzas, Tuncer Ören, and Bernard Zeigler (eds.), *Modelling and Simulation Methodology in the Artificial Intelligence Era*, North-Holland, Amsterdam.